Corpus based approach for resolving verbal polysemy in Tamil

Rajendran S and Anandkumrar M. and Soman, K.P.
Amrita Vishwa Vidyapeetham, Coimbatore 641 112

The growth in the utilization of machine readable texts in NLP tasks culminates into various corpus-based approaches. Word distinguishing techniques have been explored variously in the context of corpus-based approaches. In this paper we try to resolve verbal polysemey by making use of corpus oriented similarity based method. The notion of distributional similarity is used in NLP in a number of tasks, including areas such as word sense disambiguation (WSD), sense induction, automatic thesaurus construction, selectional preference acquisition, and semantic role labeling.

Sense Assignment

Creating sense inventory for polysemy is a task that is notoriously difficult to formalize. For polysemous verbs especially constellations of related meanings make this task even more difficult. In lexicography, "lumping and splitting" senses during dictionary construction – i.e. deciding when to describe a set of usages as a separate sense – is a well-known problem (Hanks and Pustejovsky 2005; Kilgarriff 1997). It is often resolved on an ad-hoc basis, resulting in numerous cases of "overlapping senses", i.e. instances when the same occurrence may fall under more than one sense category simultaneously. This problem has also been the subject of extensive study in lexical semantics. It addresses the following questions: When does the context select a distinct sense? When does it merely modulate the meaning? What is the regular relationship between related senses? What compositional processes are involved in sense selection? (Pustejovsky, 1995; Cruse 1995; Apresjan, 1973). A number of syntactic and semantic tests are traditionally applied for sense identification; it comprise of examining synonymous series, compatible syntactic environments, coordination tests such as cross-understanding of zeugma test (Cruse 2000). Normally a combination of factors is used as none of these tests are conclusive. There are difficulties in establishing a set of senses available to a lexical item. This is because the meaning of a polysemous verb is often determined in composition and depends to the same extent on the semantics of particular arguments as it does on the base meaning of the verb itself. A number of systematic relations often hold between different senses of a polysemous verb depending on the kind of ambiguity involved in each case; some senses are easier to distinguish than others. Treating different disambiguation factors separately would allow one to examine the contribution of each factor, as well as the success of a given algorithm in identifying the corresponding senses.

Gries (2006) analyses word senses from the perspective of cognitive linguistics on the one hand and corpus-linguistics as well as corpus-based lexicography on the other hand. While many recent cognitive linguistic approaches to polysemy have concerned themselves with polysemous words as network-like categories with many interrelated senses (with varying degrees of commitment to mental representations), corpus linguistic approaches have remained

rather agnostic as to how different word senses are related and have rather focused on distributional characteristics of different word senses. Corpus linguistic quantitative methods can provide objective empirical evidence suggesting answers to some notoriously difficult problems in cognitive linguistics. A very common problem with glossing a sense involves the situation where a sense inventory includes two senses one of which is an extension of the other. The derived sense may be related to the primary sense through metaphor. This often results in the former taking on a semantically less specific interpretations. The problem with creating glosses in this situation is that the words used may have sense distinctions parallel to the ones in the target verb being described. This leaves the annotators free to choose either sense.

The approach advocated by Kishner and Gibbs bridges the gap between cognitively oriented approaches and the linguistic paradigm in which the question of how to determine whether two uses of a particular word instantiate two different senses or not has probably received most attention, namely (corpus-based) lexicography. Organizing and formulating a dictionary entry for a word requires many decisions as to whether two citations of a word instantiate senses differing enough that the word's entry needs to be split or whether the citations instantiate senses similar enough to be lumped together. Although the lexicographer's interest in sense distinctions need not coincide with that of linguists of a more theoretical persuasion, the basic question of course remains the same. Given these questions, recent lexicographic work has arrived at the conclusion that word senses as conceived of traditionally do not exist and has therefore adopted an increasingly corpus-based approach. For example, Kilgarriff (1997: 92) argues in favor of "an alternative conception of the word sense, in which it corresponds to a cluster of citations for a word". In the simplest possible conception, "corpus citations fall into one or more distinct clusters and each of these clusters, if large enough and distinct enough from other clusters, forms a distinct word sense" (Kilgarriff 1997: 108). Hanks (2000: 208–210) argues for a focus on separate semantic components (jointly constituting a word's meaning potential), which can be weighted in terms of their frequency and predictive power for regular word uses. However, the above is only a very abstract idealization of the actual cognitive processes underlying sense identification and distinction. This and the fact that many of these processes result in apparently subjective decisions is immediately obvious once a user consults different dictionaries on the same word. Therefore, corpus-based lexicographers have begun to formulate strategies to provide a more objective foundation for resolving such issues by, for instance, identifying corpus-based traces of meaning components etc.

Resolution of polysemy in Tamil verbs

The idea that semantic similarity between words must be reflected in the similarity of habitual contexts in which words occur is fairly obvious and has been formulated in many guises (including the "distributional hypothesis" (Harris 1985), the "strong contextual hypothesis" (Miller & Charles 1991), and even the much-quoted remark from Firth, on knowing the word by the company it keeps (Firth 1957). When applied to the case of lexical ambiguity, it leads one to expect that similar senses of the same word will occur in similar contexts. However, one of the main problems with applying the idea of distributional similarity in computational tasks is that in

order to use any kind of generalization based on distributional information, one must be able to identify the sense in which a polysemous word is used in each case.

Establishing a set of senses available to a particular lexical item and (to some extent) specifying which context elements typically activate each sense forms the basis of any lexicographic endeavour. Several current resource-oriented projects undertake to formalize this procedure, utilizing different context specifications.

As we stated already we will try to resolve verbal polysemy by making use of corpus oriented similarity-based approach. In similarity-based method, which is of one of the corpus-based framework, the system uses a database, in which example sentences are manually annotated with correct word senses. Given an input, the stems search the database for the most similar example to the input. The correct sense of the word in the input is resolved by selecting the sense annotation of the retrieved example. In this paper, we apply this method of resolution of verbal polysemy, in which the similarity between two examples is computed as the weighted average of the similarity between complements governed by a target polysemous verb.

Crea' Modern Tamil dictionary lists 21senses for the verb ooTu 'run. Getting a corpus which covers up all these senses is not possible. So apart from extracting corpus from various source including web sites, we create corpus artificially for the left out senses found in the Crea. Crea has classifies the 21 senses t senses given listed under *ooTu* into three main groups:

a. Usage related to leaving a place
b. Usage related to that which moves in a fixed state
c. Usage related to expressing movementless into movementful

The list includes the following senses: moving faster than walking (as the animals which moves by placing their legs front and back) as primary meaning and moving of vehicles, moving of machines, breathing when air goes inside and outside, moving of blood, water, etc. in a particular path, spreading of the lines in palm or root of the plant, spreading of grayness in head, running of cinema, drama, etc. in theatres, moving of works, moving of (i.e. selling of) commodities, start working, and thinking to start working as secondary meanings. The following are the few examples which shows distinction between different senses.

Kuzantai kuTukuTu enRu ooTiyatu
The child was running fast
pooraaTTattin kaaraNamaaka irayil ooTavillai
'The buses did not run because of strike'
kaTikaaram nanRaaka ooTukinRatu
'the watch is running well'
Irattak kuzaaykaLil irattama ooTukinRatu
'The blood is running in the blood vessel'
piTippataRkumun tiruTan ooTiviTTaan
'the thief ran away before catching him'
taNNiir iraikkum iyantiram cariyaaka ooTavillai
'the water pumping machine is not running well'

Unkaiyil atirSTa reekai ooTukinRatu
'Yours palm has lucking lines
Talaiyil narai ooTiyiruntatu
The head has gray hair
Inta tiraippatam nuuRunaaL ooTumaa?
Will this cinema run for hundred days?
Veelaiyil ceerntu oruvaruTam ooTiviTTatu
One year has passed away after joining the job

The corpus at hand is annotated for the sense enumerated in Crea. The contextual vectors for each sense is identified and reserved as a testing sample. The testing context vector is used for identifying the correct senses of the new corpus for the target word.

Conclusion

As the field of linguistics increasingly turns to usage-based and quantitative methods, corpora can supply supporting evidence for questions answered with other methods and go beyond them in terms of both description and explanation. We have experienced from the analysis of corpus of Tamil for building lexicon for Tamil that polysemy reflected in dictionaries can be elaborated or condensed using corpus approach to polysemy.

Apresjan, Jurij D. 1973. Regular Polysemy. Linguistics 142:5-32.

Behrend, Douglas A. 1990. The Development of Verb Concepts: Children's use of verbs to label familiar and novel events. Child development 61;681-696.

Berez, Andrea L and Gries, Stefan Th. 2008. In defense of corpus-methods: A behavioral profile analysis of polysemous *get* in English. Proceedings of the 24th NWLC, 3-4 May 2008,Seattle, WA

Chomsky, Noam. 1981. Lectures on Government and Binding. Dordrecht: Foris.

Douty, David. 1979. Thematic role and argument selection. Language 67(3):547-619.

Fillmore, Charles J. and Beryl T. Sue Atkins 2000 Describing polysemy: the case of 'crawl'. In: Yael Ravin and Claudia Leacock (eds.), *Polysemy: Theoretical and Computational Approaches*, 91–110. Oxford: Oxford University Press.

Firth John R. 1957. A synopsis of linguistic theory 1930-1955. In Firth John et al. *Studies in Linguistic Analysis*, Oxford: Blackwell. 1–32.

Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: The many senses of *to run*'. In Gries, Stefan Th. And Anatol Stefanowitsch (eds.) 2006. Corpora in Cognitive linguistics: corpus-based approaches to syntax and lexsis. Berlin, New York: Mouton de Gruyter, *P. 57-99.*

Grimshaw, Jane. 1981. Argument structure. Cambridge, Massachusetts:The MIT Press.

Halliday Michael A.K. 1973. *Explorations in the Functions of Language*. London: Edward Arnold.
Harris Zelling S. (ed.) 1985. Distributional structure. In Katz Jarrod J. (ed.). *Philosophy of Linguistics*. New York: Oxford University Press. 26-47.
Hanks Patrick & James Pustejovsky 2005. A pattern dictionary for natural language processing. *Revue Francaise de Linguistique Appliquee* 10(2). 63-82.
Kilgarriff Adam 1997. I don't believe in word senses. *Computers and the Humanities* 31. 91–113.
Kishner, Jeffrey M. and Raymond W. Jr. Gibbs 1996 How *just* gets its meanings: Polysemy and context in psychological semantics. *Language and Speech* 39 (1): 19–36.
Levin, Beth.1993. English verb Classes and alternations: A preliminary investigations. Chicago: University of Chicago Press.

Levin, Beth and Malka Rapport Hovar. From lexical semantics to argument realizations. In Handbook of Morphosyntax and Argument structure, Hagit Borer (ed) Dordrecht: Kluwer Academic Publishers.

Pustejovsky, James.1995. The Generative Lexicon. Cambridge, Massachusetts; The MIT press.

Miller George & Walter G. Charles 1991. Contextual correlates of semantic similarity. Language and Cognitive Processes 6(1). 1-28.
Miriam Butt, Mariam and Geuder, Wilhelm (eds.) 1998. The Projection of Arguments: Lexical and computational factors. CLSI Publications.

Pinker, S. (1989). *Learnability and cognition: the acquisition of argument structure*. Cambridge, MA: MIT Press.
Rumshisky, Anna. 2008. Resolving polysemy in verbs: Contextualized distributional approach to argument semantics. Rivista di Linguistica 20.1 (2008), pp. 215-24

Wasow, Thomas. 1985. Postscripts. In Peter Sells (ed) Lectures on Contemporary Syntactic Theories, 193-205, Stanford, California: CSLI publications.